**CSC411 Homework 6**

**Part 1: Learning the Parameters**

**Question 1: M-Step Update Rules**

**Note:** assume symmetric Dirichlet as per instructions, with

**Final Answers:**

**Derivation:**

Objective function for reference

Note that , so Lagrange multipliers must be used for maximization. Maximizing under this constraint is equal to maximizing:

Definitions of probabilities

Applying log to each probability

Rearranging objective function:

Subbing in log probabilities

Simplifying multiplication inside summation

Deriving with respect to ,

|  |  |  |  |
| --- | --- | --- | --- |
| **Term** | **Derivative WRT:** | | |
|  |  |  |
|  | 0 |  | 0 |
|  | 0 | 0 |  |
|  | 0 | 0 |  |
|  | 0 |  | 0 |
|  | 0 | 0 |  |
|  | 0 | 0 |  |
|  |  |  | 0 |

|  |  |
| --- | --- |
| Derivative WRT: |  |
|  |  |
|  |  |
|  |  |

Maximizing

Maximizing

**Lemma 1:**

Proof:

**Question 2: Output of Part 1**

pi[0] 0.085

pi[1] 0.13

theta[0, 239] 0.642710622711

theta[3, 298] 0.465736124958

**Part 2: Posterior Inference**

**Question 1: Derivation of Posterior Distribution**

**Final Answer:**

**Derivation:**

Bayes Rule

Definitions of the above terms

Can use marginal probability to compute

Subbing in terms to original equation:

**Question 3: Part 2 Output**

R[0, 2] 2.02804472847e-27

R[1, 0] 1.11768091022e-39

P[0, 183] 0.438919126369

P[2, 628] 0.422649035891

**Part 3: Conceptual Questions**

**Question 1**

If pixel for every training sample in class , then .

Therefore, the probability of observing will always be 0 – (see definition of , equation 2 – the product will cancel into 0)

**Question 2**

This makes sense because the model has access to learned latent variables (), which it can use to compensate for the missing information.

**Question 3**

This does NOT mean that the model believes there are more 1’s than 8’s.

The higher the log probability, the more confident the model is in its prediction.

If the log probability for 1 is higher than the log probability of 8, this means that the model is more confident about judging an image to be 1 than it would be to judge it as an 8.

Intuitively this makes sense as 1’s are often near-vertical lines, whereas 8’s have features that are present in other classes:

3 looks like right half of an 8

5’s bottom looks like bottom of 8

6’s bottom looks like bottom of 8

9’s top looks like top of 8